

## **iConference 2011 Susan Dumais Keynote Transcript**

*The following is a transcript of the iConference 2011 keynote address delivered by Susan Dumais of Microsoft Research. The address took place at the Seattle Renaissance Hotel on Wednesday, February 9, 2011.*

### **Introduction by Conference Co-Chair Jonathan Grudin:**

Susan Dumais served on two NRC committees on digital government and on the board for NIST programs. She was elected to the CHI Academy. She is an ACM fellow. In 2009, she received a SIG-IR Gerard Salton Lifetime Achievement award. And I am pleased to be the first person introducing Sue at a conference who can report that she was also elected to the National Academy of Engineering on Monday, which is a great honor, so I'll do the honor for engineers.

I will say a few more words about her work and how I was exposed to it. Sue got her Ph.D. in cognitive mathematical psychology at Indiana University, and then after that she worked at Bell Labs and Bellcore until she came to Microsoft research in 1997. At Bell Labs and Bellcore she worked in a number of areas, but one of her major areas of focus was on latent semantic indexing, which is the statistical method for concept-based retrieval. I think most people here are familiar with it, and she was a leading developer of it. In 1981, 30 years ago exactly, she was an author of a paper "Verbal Disagreement: A Problem for the Information Age." So there you have the information age already underway in 1981, and Susan has been working assiduously at advancing it since then.

Sue was active in SIG-CHI from the very beginning, and she gave papers—often her papers were in the same setting that I was giving papers in, which was not a coincidence, because what inspired me to go and do HCI was a paper by Tom Landau, her colleague at Bell Labs and at Bellcore, and so for a time we were in the same area there. When I arrived at Microsoft Research one year after Sue, she was working and producing great results in an area that she felt was very important, but she was working almost alone in that area. And that area was search. At Microsoft, the view of Bill Gates and Steve Ballmer was that there will never be any money in search, but then our iConference co-sponsor Google came along and the attitude changed. And the reason why Microsoft is competing in the area of search is because of the work that Sue had been doing on that in the years leading up to that change of mindset. This enabled Microsoft to get into the game very quickly, with the groundwork that she had put in place there. Today she is a principal researcher at Microsoft Research, and she is manager of the Context, Learning, and User Experience for Search Group, and she will tell us about some of her work and interests. So thank you very much, Sue.

### **Susan Dumais keynote address:**

Thanks everyone, it's a pleasure to be here. What I would like to talk about this morning is what I am going to call temporal dynamics of information systems. And by that what I mean is simply that many of the digital resources that we interact with are highly dynamic, ever changing collections of information. Yet most of the tools that we use to examine those collections, ranging from the very basic level web protocols, search engines, and browsers, treat the world in a stateless way, as if it is a single snapshot of information, without acknowledging the fact that it is time-bearing. It was interesting to hear Karen Fisher in her opening remarks talk about how the iConference has really evolved over the years. Yet that is very hard to tell looking at a single snapshot of information. What I am going to talk about today is work that I have done in collaboration with lots of colleagues mentioned here to do two things. One is to try to characterize and understand how content changes over time, how people's interaction with that content changes, and then to develop tools and methods for supporting people in understanding the evolution of information and ideas over time. What I am going to do is highlight at a very quick level some of the differences that make me excited about digital dynamics.

The first is that there are many differences between physical and digital libraries, so this is a picture of the Seattle Public Library and a conference homepage. But what I am going to focus on today is that change is everywhere in digital collections. And to be fair, there is some change in traditional paper libraries, there are certainly multiple editions of books published, there is an acquisitions process, and people borrow and interact with the information. But change, I will argue, is really pervasive in digital collections, and part of our job I think is to help people appreciate those dynamics, to understand how ideas evolve and to process this all. So I am just going to mention a handful of ways in which digital content is constantly evolving over time. First, new documents appear over time, this is a small screenshot of Twitter on New Year's Eve in Japan, where they reached about 7,000 tweets a second. Early in January, Twitter hit something like 110 million tweets a day. Facebook probably gets a billion status updates a day. These real-time social systems are highly dynamic. Queries and query volume change over time. Information retrieval people like to think of a static test set, and static collections, but query volumes are changing constantly—they are not identically and independently distributed over time. To give you an example of a query of Lalapalooza, you can see a couple peaks there. One is when they announced who is going to appear at Lalapalooza, the other is when tickets go on sale, and there is a third peak not represented there about when the event actually happened. Similarly, conferences have multiple peaks in the query volume: one when papers are due, one when the notifications of registrations come out, and the third when the actual conference happens. And what should I serve people, what should I return to people asking questions about the same event at different points in time? Documents change over time, even a page such as the iConference home page or my personal home page, which we will see later, look somewhat static, but they actually do change a lot if you look deeper inside. If I ask a query, say U.S. Open 2010 or maybe 2011, what I want in May is typically the golf tournament; what I want in September is a tennis tournament. A query like Hurricane Earl, if you were on the East Coast in

September this past year, it meant the tropical storm that was streaming up the East Coast. If you ask for it now, you probably mean the much bigger hurricane Earl which happened in 1998. And so what people mean by a query and what is relevant to that query, what we want to serve them changes over time. In addition, it is important that user interaction with that information changes over time: Tags are annotated, “likes” are added, anchor text is created, the interrelationships among documents, among people—those are all constantly changing over time, and so change is really everywhere in digital information systems. Yet we are not doing very much about it.

Here is a schematic I am going to use throughout the talk, and it represents simply the evolution of a page, and this page is the Microsoft Research home page from 1996 to this past year. The content on that page gets refreshed often. Similarly, people visit that page and revisit that page over time. But the tools we use to access that, whether it is a search engine or a browser show a single static snapshot—the browser shows you what is there when you visit; the search engine shows you what is there when the page was crawled. And what I want to do is talk a little bit about how we can look at information that has existed in the past, in the immediate past, and also in the longer term past. This schematic about the evolution of the Microsoft Research homepage tells you a lot about that organization, about its cadence, about its livelihood. And right now we tend to drop all of those on the floor.

Digital dynamics are really easy to capture. Here is the Microsoft Research homepage on Monday morning. This is what it was like in 1997. You can't make this stuff up. This is a serious professional site with a computer imploding and smoke coming out of people's ears. So the look has changed a little bit. This was the UW iSchool's homepage about a week ago, nice Husky purple and gold, a very professionally done site. Nice advertisement for the iConference, nice skyline of Seattle and so on. This is what it looked like in 2005, almost as if it was the invasion of the North Carolina Tar Heels—it turned from purple to blue. But it has all sorts of really fun information, one of them it had a grant that Harry Bruce and William Jones had just gotten from the National Science Foundation on personal information management. Cheryl Metoyer had a nice grant that was featured there. Lee Dirks had come to give a talk. So that was what it was like in 2005. This is what it was like in about 2000, it reverted back to the purple and gold. You can see that the Mary Gates hall was just being dedicated, Mike Eisenberg was Director not Dean of the school at the time. Both of those pages, Microsoft Research homepage and the UW iSchool page have very visible changed.

Here is my homepage today. Here is my homepage 13 years ago. There have been a lot of changes. No really, I assure you it is just an issue of being below the fold, there is a lot that is going on there. It doesn't look like a lot happened, but in fact if you look at it more closely all sorts of things happened. I have changed titles, I have changed groups, I said I was going to do a lot of work, and now I am saying what I am working on. The name of Bellcore changed from Bellcore to Telcordia. And all of those changes, not salient visually and are completely hidden if you just flip back

and forth between these versions, and so as I am going to talk about our tools that allow us to see some of those dynamics.

What I am going to do today is start by trying to characterize how content changes over time and how people revisit and re-find information over time, and then, more importantly, I am going to talk about how we can take that understanding and leverage it to improve how information is retrieved, how people understand and work with information. I am going to give examples from my own work on search and browser support, but the ideas as we will see are obviously much more general. I am going to start by talking about two older projects on desktop search and personalized news, each of which dates back about seven or eight years, and then I will talk about some more recent work on web analysis. I am going to start with a project that was called, very technically, *Stuff I've Seen*, and this went public in 2003 so it probably started about a decade ago when I was really frustrated by the fact that it was much easier to find stuff on the web than on my desktop. And there is Ed Cutrell [in the audience] shaking his head—he was one of the early collaborators in this work. Part of the reason it was hard to find information on the desktop was that although Windows actually shipped with desktop indexing, it was turned off by default. But more importantly there were many silos of information: there was what I had on my desktop, there was what I had in the browser, what I saw on the web, what I took in journal notes, what I took on OneNote. There were lots of different sources of information, each with their own information organization and search capabilities.

What we decided to do with *Stuff I've Seen* is build unified access to that information. We weren't going to unify the storage, but we wanted to provide unified access to the distributed content. What is interesting is you see the same effect happening today where we have social media of all different kinds, and it is represented in a very distributed and sort of piecewise fashion. So we built unified access to mail, files, web pages, RSS feeds, and IM. We indexed the full content of this information as well as metadata, and we provided very fast and flexible search. We were interested also in providing support for re-finding and not just finding and serving information in the first place. On the desktop especially you want to re-find information you have seen before. The work on *Stuff I've Seen* influenced in many ways the latest versions of desktop search, but what I want to focus on today is what we did on *Stuff I've Seen*.

How many of you use desktop search? Okay. Oh, very nice. So you understand the scenario. Before I gave a recent talk on *Stuff I've Seen* I sent an email to colleagues asking them to describe the last thing they looked for on desktop search, where it was initiated from, and then what the query was. Here's one of somebody looking for a recent email from Fedor that contained a link to a new demo. Here's one of somebody looking for the pdf of a SIGIR paper on context and ranking—not sure if it used those words, but somebody, don't really remember who, sent me about a month ago. This was mine—I knew some things very well and others not so well. Here is a meeting invite for the last intern handoff and C# program that I wrote a

long time ago. What is interesting about these searches is that they contain a lot of metadata, so every single one of them contains a reference to *when* something happened as well as lots of other information that the searcher remembered about it. So we thought that this kind of desktop searching interface that we developed would be really helpful at helping people access this kind of information. This is a fairly standard and sort of ugly-looking rich list view that has lists of items and you can sort by any of the columns and sort by title, date, author, and so on, and a bunch of filters for helping to slice and dice the information.

We studied *Stuff I've Seen*. We actually built it in a way that initially, hundreds folks sent data, and what I want to talk about here is from the initial 250 people, but over the years about 3-to-4,000 people wound up using it. We studied this in many different ways. We had free-form feedback. If you deploy anything to 3,000 people at Microsoft you get a lot of unsolicited feedback. We studied it with questionnaires, with structured interviews; we collected some log data, actually pretty minimal log data. We collected basically how people interacted with the application so we knew the rank of an item they clicked on, the date that it was accessed, but nothing actually about its content. We also did some in situ experiments, so different people who signed up to download to *Stuff I've Seen* got different default settings, and then we also used lab studies to study more detailed and complex effects, the “why” as well as the what—a richer data about what people do.

People's personal stores, at the time that we did this, this was about seven or eight years ago, ranged from 5,000 to 1.5 million items. Just as a point, my laptop here contains 100,000 items and I add new items at the rate of about 1,000 per week. The other interesting thing about desktop search is that it is not like web search. One of the biggest differences lies in the fact that people know a lot of metadata about what they are looking for, and they want to be able to express that. People are extremely important: almost 30 percent of the queries contain somebody's name or an email alias. Date is exceedingly important, as I alluded to earlier. Date was by far the most common sort order for the results even when we gave people a different alternative. We had this fancy best match algorithm like Okapi's BM25 which used all these factors, and more often than not even people who had that ordering would switch to date order. I have done user studies for decades, and when people switch away from an alternative, it tells you something about what a good starting point is—it doesn't tell you that date is always the best, but were I to pick a single default (for desktop search) that is what I would pick. And the reason for that is that there are very few searches on the desktop for the “best” matching item. Now, I got mail from Harry last week, it doesn't really make sense to look for the best mail from Harry. When I was looking to find that mail I wanted to find the mail about when he told me this talk was going to be and that it was changed. I don't want the best mail from Harry no matter how fancy the ranking algorithm is; I want to find the one with the particular attributes that I remember. People remember a lot of different criteria about their own desktop information, and we really need to support that kind of flexible access. We also looked at the age of articles that people retrieve from the desktop. Five percent of them are things that people first saw today, 20 percent

during the last week, 40 percent during the last month. But that also means that over 50 percent of the items that people were retrieving were more than a month old, and that suggested to us that we really needed to support episodic access to memory.

I will now describe the system that we built to do that. In cognitive psychology there is a notion made popular by Endel Tulving that memory is organized into episodes; it is not organized by absolute dates. Right now I know that I am speaking at the iConference at 9:30 in the morning on February 9th, and ten years from now I will probably remember it as relative to some important event in our lives. My husband and I are taking off for our anniversary next week. I will probably remember the occurrence of this relative to one of those lifetime events that is either historical or a personal event. What we were interested in doing was seeing whether we could use these kinds of semantic and cognitive landmarks to facilitate search, and we did so by developing a kind of timeline interface that was augmented by memorable landmarks. So here is a set of results in a not-so-fancy looking interface—type in a query, you get a list of results ordered by time. On the far left we also showed the distribution of those items over time. And one thing that you can see immediately is that the distribution of results is not uniform over time. This is a query for earthquakes, so it happens to be good that it does not occur regularly over time. You can see that there is a cluster of events here, there is another cluster here; that in and of itself tells you something and can help you hone in on relevant results just by knowing whether it happened during a burst of activity or happened during a lull. We also added what we call the memory backbone, and what this is are events that we thought would be salient to users that are only related to the results by time, so these events really have nothing to do with the query, they just happened at roughly the same time as the query results. This item up here, I see two people in the audience know it very well—a the CHI 2001 meeting happened in Seattle shortly after the earthquake. Nick, Colleen, Mary-Jo, Mark Altom and I went out wine tasting after, and this is a picture from that. So I can tell you exactly when that was; it is a very salient cognitively for me, although probably for nobody else in the audience; these are my personal timelines. One of the things we did was to develop computational models to identify what would be memorable events, and the interesting thing is that we can do this with very high accuracy. So for example, for modeling landmark meetings, we had people label up to I think ten years of their calendar events as to whether they were memorable or not. Jonathan was actually one of the participants in this—he probably remembers that labeling episode as still memorable today because it was something tedious if nothing else. For me landmark events are characterized by being atypical, so they are different in location, in terms of the organizer or attendees, those make landmark events. My role in the meeting is also relevant to whether it is a landmark or not. If I am the organizer, it is more likely to be a landmark event. The important part is that we can develop computational models to identify important events; these are calendar events, but we can do it for a variety of other kinds of information, and then you could anchor search results using those important events.

Eric Horvitz and Paul Koch have developed a system that is sort of memory landmarks on steroids; it is called Life Browser. It shows these landmark events of all sorts—images, desktop search activity, appointments, events, whiteboard capture, whatever you interact with on your desktop. So whether so you are exploring a file system, doing a search and looking at results or whatever, it is a really fascinating system to live with because it is anchoring everything in the context of other ongoing activity.

Okay, let's switch gears a little bit and move away from desktop search to an analysis of news that we did six or seven years ago now. News is a really interesting example of a stream of information that that evolves over time, but it is really hard to consider news as a stream of information. If you go to a traditional newspaper or even a newspaper's website, what you get is what they choose to publish, when they choose to publish it. If you go to a news aggregator, you might see that the iConference happening in Seattle today had 20 articles associated with it at 7 a.m. and by noon it will have 300, but that doesn't give you very much information about what has changed between 7 a.m. and noon. What we have attempted to do in this system is provide a personalized news stream that looked at the novelty of information relative to what you already knew about it the event. So to do that we identified clusters of related articles, much as you see in a news aggregator site, we then tried to characterize what a user knows about an event already, and then we compute the novelty of any new article relative to what you already know about the event. So now when I go back to one of the news aggregators, with this system, it would say here are three things that are new relative to what you already know, not 500 that are recaps of what you already know. We use a simple measure, the KL divergence between the text representation of an article and my current knowledge of the event; and then we can use that novelty score to drive what to show somebody, when to show it, and how to show it to them.

I can give you one example of *NewsJunkie* in action. This is an event that was happening as we were working on it; some of you may have heard about it. It is actually a very sad but sort of amazing story that evolved over almost a decade. This is an accounting of Brian Wells, who was a pizza delivery man in Erie, Pennsylvania, and he delivered pizza, then robbed a bank, the police arrested him, and he said to contact the bomb squad, because people had put a bomb on his neck. And while they were waiting for the bomb squad to arrive, there really was a bomb on his neck that blew up and he died. So this story evolved in crazy ways for the next decade. So here is how we analyze it with *NewsJunkie*. These are articles over time, and this is the novelty score given what I already know about this event. You can see several peaks (of novelty), so something here interesting happened that is not what I already knew, something happened here, something here, and something here. In this particular case, there were a lot of news reports, his friends talked about the guy, they then found that the cane he used was actually a gun, they were looking for two people who had suspicious activity at the time, there then was a copycat case in Missouri, and this evolved. This was just during the first week, the event later evolved to finding people who lived next to where the pizza was delivered and

finding bodies in coolers, and it was just an insane story that ended actually just this past November when Marjorie Dhiel-Armstrong was convicted of being one of the people involved in this, and it was just reported two weeks ago on NPR. But this is a story that has evolved maybe more than most evolve over time, but we think that *NewsJunkie* provides a really interesting way to stay in touch with evolving events based on what I already know.

So those are two older projects having to do with desktop search and news. I am now going to switch gears and talk about characterizing change on the web. If you look at my webpage or the iConference web page, those seem like relatively static entities, but as I mentioned before they actually do change a lot over time. We studied that by crawling lots of web pages very regularly. We looked at two sets of web pages that I am going to talk about today. One was 55,000 pages that people revisited, so we sampled these pages because they were pages that people revisited with different periodicities. And we crawled those pages every hour for a year-and-a-half or maybe two years. We also took a somewhat more random sample of web pages. We took six million pages that were returned by a search engine for a large number of test queries, and we crawled those every two days for six months. So we have these two large collections of pages crawled periodically over time, and we can analyze how the content of those pages changes. There are a number of ways in which we looked at page change. I will talk about some really high level summary metrics, I will talk about a new analysis method that we can all call change curve, and then I will talk about within page changes. All of that will then give us insights about how these objects are evolving over time. So just some change factoids: of those six million pages we crawled, about 33 percent of them changed over the course of a month. Of the visited web pages, these are pages that people actually visit, not just some random page on the web, about 66 percent of those change every month. So things that people are visiting change or people visit things that change, we don't know which. And, of those changing pages, 63 percent of them change every hour. So there is a lot of interesting dynamics going on in content that we are literally blind to because of the way that browsers operate and the way search engines operate. For the amount of change, the average Dice coefficient is about 20 percent—20 percent of content is new every time it changes, and the average time between changes is about five days. Actually .edu and .gov pages don't change very much and they don't change very often. Popular pages change much more frequently but again not by much. I think the .net pages change with some intermediate frequency and they changed a lot. So there are really interesting things that you can understand about the world sliced in different ways by looking at these change measures.

We also developed a somewhat more refined measure to look at change. We start with a web page, so this is the WSDM conference home page, and see how much it changes at different points in time. We will look at the WSDM home page on day one, on day two, day three, day four. And if you plot the overlap between pages as a function of time, not surprisingly it decreases. So this version of the page is less similar to the first version, than to those that occurred after. It is not so surprising



that the similarity between pages changes over time, but what we were interested to find is that most pages have this kind of bilinear function and this inflection we called the knot point. The slope of this function tells us how quickly stuff disappears off the page, and the asymptote tells us how much stuff is really consistent over time on a page. So the X-axis is going to tell us how fast information is disappearing from a page and the Y-axis is going to tell us what the steady state similarity of that page over time is. These are change curves for two pages, the top one is Craigslist Los Angeles, and that asymptote is after about ten hours, and the asymptote is at the level of about 40 percent overlap, so within ten hours 40 percent of the stuff on Craigslist is no longer there, and that is true going out further over time. On allrecipes.com the half-life of that information is about 260 hours that is about 10 days, so it drops off much more slowly, content there changes much more slowly, and its asymptotic level of similarity with the original pages is more like 85 percent or 75 percent. So that page changes less rapidly and is more similar to itself over time.

We also were interested in understanding the details of within-page change. Here is a page from allrecipes.com the homepage. We wanted to understand which words are changing and how are they changing. We did that by measuring two things: one is a divergence from the norm, this is sort of like an IDF weight for the term. So imagine you have the page allrecipes.com, you want to find out which words on allrecipes.com are different from the rest of your collection, and for allrecipes.com the words that are different are things like "cookbook," "salad," "cheese," "ingredients," and "barbecues." We are also interested in the staying power of a word on a page. This is a kind of funny graph that shows time on the X-axis and words on the Y-axis. You can see that there are some words like "cooks" and "recipe" that are there in September and are there every single hour the page was crawled for November and December. Others, like "frightfully," appear in October and disappear pretty quickly. "Sugared yams" and "merrymaking" appear in December and disappear pretty quickly. On this page the words that have both high divergence from the norm and that are unique to allrecipes.com and that are there over and over again are things like "cookbooks" and "ingredients." In some sense those words are characteristic of the page and others are more transient information about the page. These are just different term longevity graphs for different pages. You can see that allrecipes.com is pretty well done editorially. There are bursts of activity maybe weekly or daily. Craigslist is changing much more smoothly. And who knows what is happening on Best Buy other than it is a little bit of a mystery. So that is how content changes.

I am quickly, in a couple slides, going to talk about how people revisit information over time, and to do this we analyze logs. We looked at browser toolbar logs to understand how people revisit content over time, and then query logs to look at how people re-find information over time. We had a survey to really understand users' intents behind revisitation and re-finding. So let's do a little audience participation here. So what was the last webpage you visited and why did you visit or revisit it? Liz?

Liz: Google News.

And had you been there before or was it a new discovery?

Liz: Every day.

Okay. Jonathan?

Jonathan: New York Times.

I visited the iConference website before. I tried to visit the iConference website before I came. So revisited a page I had previously visited. People do visit new pages, as well. Yeah, Mike?

Mike: I visited [a Seattle-based travel site].

Okay that was a new page for you. People often go back to information they have been to before with the intent of discovering new information like the news. I was trying to discover information that I knew existed on the iConference home page and Mike was looking at a new page to discover information about what users like to do in Seattle. So people revisit for lots of reasons. Here we have some summary metrics of re-visitation. For 60-to-80 percent of the web pages you visit, you have visited the page before. As I have said, people do so for a lot of reasons. This suggests that although we are very focused in the information retrieval community on helping people find new information, we should be equally excited and provide support for people revisiting information and highlighting what is new. So by analogy to the change curves I showed you earlier we also have re-visitation curves, and here what we do is start with a page and look at the interval between successive visits, and you count how often that happens. For this page particular page, the WSDM conference site, the peak re-visitation interval is something intermediate—you revisit after long periods but not very often, more frequently revisits are for short-lived information. You can now compare these re-visitations to change curves, and I am not going to do that in detail, but I am just going to highlight there are lots of relationships between change and re-visitation. People revisit because they are interested in change, so I might want to check a stock price, or Liz might monitor the news pages. I might actually go to a site to affect change; I might want to cause change. I might log into a mail site or editing a site. And change may actually be unimportant. I may go to the iConference page not to pursue new information but to find what I have seen before, and in these cases actually changing a page can interfere with your find. Re-finding, it turns out, is very important in search. People repeat queries a lot. I have repeated the query iConference 2011 probably 50 times over the last month, maybe more. Thirty-three percent of the queries that people issue on the web are queries they have issued before. People revisit the same pages. Thirty-nine percent of the things you click on in response to a search query are things you have clicked on before. Sometimes it is for the same query; sometimes it is for different queries. I have discovered recently that in Bing if I just type in iConference I get the 2011 one; the minute I discovered that I stopped typing 2011 after it, and so now I just type iConference rather than the whole thing. But people do a lot of repetitive behavior, and we should help them, I think, in understanding that change. There is a big opportunity here; almost over 40 percent of people's behavior involves repeat clicks to web pages or repeat queries. And I think that is as important as focusing on information discovery.

I will now turn to a couple of new systems that we have developed to support people in re-finding old information. One is a browser plug-in called Diff-IE [ <http://research.microsoft.com/en-us/projects/diffie/> ], and the second is some temporal retrieval models which I will get to. Diff-IE is a browser plug-in that highlights the changes to a webpage since you last visited. You don't have that ugly Word revision control marks—cross out, red, highlight, or side notes. This is a very simple idea; it is actually available on Microsoft Research [ <http://research.microsoft.com/en-us/projects/diffie/default.aspx> ] starting last month for folks to try and download. There are four interesting features of that that I want to highlight. The first is that it is always on. Unlike an RSS feed or other sort of notification system, you don't have to subscribe to it; it is there all the time. And as I will show you later, I think some of the beauties of this are to discover to look at pages with a different lens than you have before even if you don't expect it to be changing. So it is always there, it is in situ, I don't have to change to a new application to view it. As I browse a web page it is there. It allows you to see what is new and will highlight in yellow, although you can change that if yellow is not your favorite color. And, finally it shows what is new to me; it is what is new relative to my last visit. Lots of new sites do things like put “new breaking news” and that is terrific, but that is from their point of view. If I just visited the web site 30 seconds ago it is not breaking news to me, I have seen it. And so this is new to me. And I will show you some screenshots of it. This was a screenshot of the *New York Times* on Saturday morning while I was working on this talks. I had already read about the elections that would be held in Egypt, and even though the *New York Times* thought that it was new and breaking information, I had just seen it a few minutes ago, and so that was not highlighted when I revisited. I use it to monitor, so this is my Twitter account, and I can see immediately who has posted on it. Some of them are not interesting, like Loren Terveen posted 22 hours ago; last time it was probably 10 hours ago; but others are new content.

There are also a lot of cases where you have unexpected changes in content. I went to the iConference homepage, the keynote page recently, and there was a long introduction to the keynote speakers, but more importantly when I went to the program it highlighted that my keynote would be held on Wednesday instead of Thursday, so that was something that was relatively fun. It also I think really helps you understand page dynamics, so here is a set of search results for the query “Jaime Teevan,” a colleague of mine. And you would be surprised to see how often search results change. Whenever I talk about personalization, people think, oh, what happens if the results I saw are not there the next time I try. Get over it. They are not there the next time you try, a lot of them aren't. But this is also a really interesting results page, where I type in Jaime's name, the first two results are actually the same, the third was a DBLP entry, and there is a new DBLP from 2010. I can see that she also has now more connections on LinkedIn and that she had won the Technology Review 35 Award for new researchers. So these are all things that are relevant to Jaime but I can now easily see what is new since my last visit.

I think it really does help put on a new lens to help you appreciate how much the world that we live in is really changing in important ways. A lot of the changes that you go to consume are expected for news and monitoring activities, but a lot are really unexpected yet truly delightful. I recently went to somebody's homepage and saw that they won a teaching award. There is no way I would have noticed that in the midst of everything else that was there. I was focused on trying to find a boring paper or something, but they won a teaching award, which is probably more important. Diff-IE I think is really an interesting way to help people appreciate, understand, and interpret the change in their digital environment. We studied Diff-IE in lots of ways. It has been deployed internally at Microsoft by 3,000 people and externally by 11,000 so far. So join up and send us your feedback. We studied it, again, in lots of ways. We had feedback buttons on the toolbar. If you found a delightful experience you could click on this and it would send us a smiley face describing what the page was like. If you were not so satisfied you could give us a frowny face, and we paid a lot of attention to that. We did a survey both before and after people had used it a month. We logged the URLs that were visited, actually hashes of the URLs that were visited, as well as the amount of change, so we don't know what you were visiting, all we know is it is the same thing you visited yesterday that it had changed either a little or a lot.

Then we did some experimenting. There were several really interesting things. After a month of using Diff-IE, people's perception of how often they revisit pages hadn't changed very much, but their actual re-visitations had increased by 15 percent. So after using Diff-IE for a month, people were revisiting more pages than they had in the past. So one question was why they were re-visiting more pages. Their perception of change on the web has actually increased by using Diff-IE, and the amount of change they see increases. So a larger proportion of their visits have changed, and they change by more, and we believe this is happening because Diff-IE provides you a way to make sense of that changing world. It is fun to go back to news pages more often than you would, because you don't have to tediously go through and decide whether you have read it already. So I think that is really exciting not to have just this fun experience but to help influence the way in which people consume the digital dynamics with Diff-IE.

There are lots of other examples of people trying to provide people with a more dynamic interface to information on the web, Eytan Adar, developed a lovely system called Zoetrope that allows you to playback a webpage over time. There is a fun system at CHI this year called Diffamation that shows you a visual transformation, a smooth transformation, from one version of Wikipedia page to another, so you go from one and smoothly zoom into another, and people work on temporal summaries and snippets.

There is also a lot of interaction with content that changes. I think still one of the most interesting prototypes in this area is something that William Hill and James Hollan and others built in 1992, almost 20 years ago, called Edit Wear and Read Wear. And the idea is that people leave digital footprints on information. You can

show in a program or on a web page where people are engaging most often on that page. And it is again sort of like Diff-IE a very non-intrusive but interesting view about where the activity is. So in addition to Diff-IE, which I think is a great user experience, we have also tried to leverage dynamics to improve core retrieval algorithms.

The basic idea here is that current retrieval algorithms look only at a single snapshot of a page. The indices of all web search engines, your desktop index, whatever it is, is index what that page looked like when it was crawled. But web pages are changing all the time, and the interesting question we posed was whether we can leverage the patterns of change for retrieval. We know that pages change at different rates. What we try to do is use change priors instead of link structure to set in a non-uniform prior on pages. We also know that terms have different longevity on a page, the slide identifies what the salient and important terms on the page are. To use temporal information for retrieval, we use a language modeling approach to ranking in which we identify probability a document is relative to a query and this is a product of a prior on a document as well as the probability that the query is generated by the document, so we are going to use information about how frequently pages change to set non-uniform priors on objects and we are going to use term longevity to provide a more interesting term weights. It turns out page relevance is related to page change. We are lucky enough to have a large set of data that has human relevance judgments associated with it. For web queries, we have hundreds of pages rated on a five-point scale as to whether they were perfect, excellent, good, fair, or bad. And it turns out that the rate of change of a page is highly correlated with whether it is a good page or not. It is not so surprising in retrospect. Curated pages are pages that editors take a lot of time and energy to maintain and provide new content. So we believe that we can use the frequency and patterns of page change to set some prior on a page and so we can set document priors based on change in the same way that you would use page rank or re-visitation data to set priors on documents by this. It turns out that words on a page also vary over time, as I have shown you, and here what we are going to do is represent a document is not just a bag-of-words but three bags-of-words, as words that evolve at different rates. There are some words that are fundamentally about the page that are there every single time, there are some that come and go, and there are some that are just very briefly there. So what we are going to do is use this model and estimate these parameters, and use this new term longevity weighted model to improve the query.

We crawled 2.5 million documents every week for ten weeks. We were also lucky enough to have 2,000 queries that were sort of navigational queries, like iConference 2010 goes to the iConference 2010 webpage. The reason we picked navigational queries is because we assumed, I think largely rightly, that the pages relevant to navigational queries are pretty consistent over time. And I will tell you later why that may not always be true, but what I can say right now is that navigational queries like *New York Times* leading to the *New York Times* homepage, iConference 2010 leading to the iConference 2010 homepage are going to have the

same relevant documents over time. And this is largely a pragmatic issue, because we have relevance judgments that were not temporally based. When you do retrieval using this dynamic model that looks at terms of different longevities, it improves retrieval by five percent. If you add the change prior part it improves retrieval more, and if you combine them you even get better retrieval of these navigational pages. So it is actually pretty exciting that understanding how the words got to be on the page, understanding how they evolve on a page can really help us improve ranking.

This initial evaluation, as I just mentioned, focused on navigational queries and we assumed that the page that was relevant to the query was static over time. But there are lots of other cases, as I noted in the introduction, so that U.S. Open means different things at different points in time. The query iConference means 2011 now, but shortly it will mean 2012, so what's relevant may change very quickly. We are now doing an evaluation in which we are collecting explicit relevance judgments every week, and at the same time crawling those pages looking at interaction data and query frequency, and we are developing both improved temporal IR models and temporal snippets.

Here is sort of a fun example. For those of you who know me, I am kind of a basketball junkie. There is this thing called March Madness, which is the NCAA college basketball tournament that happens in March. And what I have listed here is judgments over time—we got judgments last year around this time for hundreds of pages relative to March Madness (and about 100 other queries), and what we wanted to look at was the relevance of those pages over time. During the tournament, the page that is most relevant is the NCAA.com page or CBS sports page, because they have the latest up-to-the-minute scores. You could even see how your team is doing. They feature the brackets; which is really where you want to go if you want to understand who is playing who and when. A general page, like this Mahalo page (but Wikipedia pages also similar ratings), are very good pages, but they are a little bit below these really dynamic ones. And this also illustrates an interesting temporal dimension, this bottom page is not rated very well at all because it is NCAA March Madness 2008. That was really important in 2008, but it is not so important in 2010. So this is what happens during a tournament. If you look after the tournament, the general web pages that talk about March Madness are still very relevant, but the sports pages that feature the up-to-the-minute scores are now no more relevant than the 2008 thing. They are off featuring tennis, or lacrosse, whatever it is in May. Search engines, even a desktop search engine or enterprise search engines, need to worry about what it is that people mean, and I think that understanding the temporal dynamics is really critical in doing that, because what people mean by March Madness, which seems like a very simple sort of concrete event, changes pretty rapidly over time.

There are other examples of work that has gone on in building information systems that support temporal dynamics. There is work on query dynamics, document dynamics, and temporal retrieval models. There is also some really interesting work

on trying to extract temporal entities from webpages. So mail is pretty easy; we know when it was sent. But when “it” is a webpage—what is a webpage about, what time period is it discussing, when was it written, when was it updated—all that kind of temporal metadata is completely thrown on the floor. There is also some interesting work that Van de Somple and his colleagues have been doing on http protocol extensions that allow you to save multiple versions of a page and get to that through the same URI. Right now you have to be sort of a wizard of URL syntax to try to get back to crazy home pages. It took me a long time to find Harry and Mike’s pages from 2000.

What I have tried to do today is convey some of my excitement and some things about temporal dynamics of information systems. Web content is constantly changing. How people visit and re-find information is changing. We can understand something about what people mean by a query by understanding the relationship between change and re-visitation. And I have also tried to provide examples from desktop search, from news, and from the web about how we can build tools like Diff-IE and temporal retrieval models to really help people understand and appreciate the highly dynamic and constantly evolving environments in which they live. As I have said before, temporal dynamics are pervasive; I think it is a tremendous area for opportunity. Just yesterday I was looking at the UW iSchool calendar. There is a student who did a thesis defense on saving and re-accessing personal information from social media over time. Bob Allen gave a talk on historian workbenches. Actually both of those feature very importantly an explicit representation and reasoning of time.

Time influences almost every aspect of information systems, from the core level of protocols, crawling and indexing the past. How do you do efficient caching of evolving things over time? There are really fundamental questions. Document representation—how do you generate metadata about a page? How do you try to index it? How do you do extraction of information over time—the chair of the iConference is different this year than it will be next year. But all of this stuff that we do if we have a static collection I think completely changes when you have a dynamic collection. User experience and evaluation are also really hard in a dynamic collections. But I do think that by better supporting the temporal dynamics of information—in order to better support the temporal dynamics of information it is going to require the explicit extraction of temporal metadata, a real focus on digital preservation. We may choose to live in an ephemeral web world, but that should be a choice. It shouldn’t be a byproduct of the tools that we are given right now. And this kind of technology I think will really enable us to have a much richer understanding of the evolution of important ideas and relationships and trends over time.

I have spent the last hour talking about time, but it is only one important aspect of context. I started to mention I had a group called the Context, Learning, and User Experience for Search group, that is called Sue’s CLUES, and I do think that context, I am totally excited about time because most other people are ignoring it, but I think

there are many other aspects of context that are important: where I am when I ask a query is important, who I am, what tasks I am doing? When I talk to an information retrieval audience I often start by putting up these three boxes. When most people think about search, they think about the little rectangle that you type in your 2.1 words, they think about the ranked list of results that come back, and if you are an IR nerd like me you worry about the black box in the middle. But really think outside the boxes. Folks here I think are converts. Searches come from real users— understanding where the queries are coming from, the rich interrelationships and evolving interrelationships among documents, as well as what the heck people are trying to do when they search is really important in developing great search systems. And I love search, as Jonathan mentioned I spent 30 years of my life doing it, but I don't get up in the morning and say, I've got ten minutes to kill, let me search. I am searching to do something. Understanding what it is that I am trying to do should inform both the results I get back as well as the way in which they are presented.

Thank you for your attention. You can find more information about me and the papers that I have mentioned in the upper right hand corner of most of the slides. And give Diff-IE [ [href="http://research.microsoft.com/en-us/projects/diffie/](http://research.microsoft.com/en-us/projects/diffie/) ] a shot. If you type in Diff-IE on Bing it is the number one result. And while you are waiting to download Diff-IE and use it, I will show you pages from Harry and Mike over the last ten years. One of them has a bigger smile and no tie on this year. That will let you know quickly who the current Dean is.